ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446718

TOPOLOGY-REGULARIZED SELF-KNOWLEDGE DISTILLATION FOR TRANSDUCTIVE-INDUCTIVE LEARNING OF BRAIN DISORDER DIAGNOSIS

Yanwu Yang[1,2]*,Xutao Guo*[1,2]*, Guoqing Cai*[1] *Chenfei Ye*[2,3]*, Ting Ma*[1,2*]

[1]The School of Electronics and Information Engineering,
Harbin Institute of Technology at Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]International Research Institute for Artificial Intelligence,
Harbin Institute of Technology at Shenzhen, Shenzhen, China

ABSTRACT

Recent advancements in fMRI-based brain disorder diagnosis have shown that graph neural networks (GNNs) have been state-of-the-art methods for brain network analysis. Among them, transductive and inductive learning can be exploited by GNN. Transductive graphs, such as population graphs, take each subject as a node and use the node classification task for diagnosis. This line of work suffers from high computational costs and poor scalability to unseen data. Inductive methods, on the other hand, only consider labeled data and may suffer from overfitting and poor generalization when training with insufficient samples. To address these limitations, we propose a unified transductive-inductive network to study the properties of both transductive and inductive learning frameworks. Our approach is implemented in a self-knowledge distillation architecture, where transductive predictions are distilled from a transductive population graph network into an inductive network as a self-supervised regularization term. To preserve the topological properties within transductive graphs, i.e., inter-node similarity, we propose a topology-regularized self-knowledge distillation (Topo-KD) approach to regularize the student model's learning. Evaluations on the ADNI dataset demonstrate the superiority of the approach in performance and scalability.

Index Terms— Graph neural network, Neuroimage, Transductive learning, Inductive learning

1. INTRODUCTION

In recent years, the functional connectome of the brain, derived from functional MRI, has become an essential foundation for brain disorder diagnosis [1, 2]. The connectome paradigm has shifted the connectome-wide association studies (CWAS) from multi-variant analysis [3, 4] to deep learning. In contrast to convolutional neural networks (CNNs), graph neural networks (GNNs) can capture the inter-node dependencies and their interactions. GNNs have become powerful tools for brain networks and could lead to significant advances in brain disorder diagnosis and treatment [5, 6].

One type of GNN architecture is to represent a population of participants as a single graph, where participants are interconnected based on connectome and phenotypic similarities [7, 8]. These approaches utilize the transductive framework and can leverage features and inter-node similarities to predict the test set via node classification. Such approaches are particularly useful and natural when only a small amount of supervised data is available [9]. However, they are computationally expensive and not scalable to new data. Another GNN architecture involves tackling each brain as a graph, with each brain region represented as a node, e.g. BrainGNN [10]. These methods utilize the inductive framework, which is more suitable for applications with large-scale data because they can efficiently and effectively deal with new data [11]. However, most inductive learning models only consider labeled data and suffer from over-fitting and poor generalizability when training with insufficient training samples, especially for hard-to-collect brain connectome data [9].

In this study, we propose a unified transductive-inductive architecture that utilizes the advantages of transductive population graphs and inductive frameworks. Our approach is a self-supervised learning network that utilizes the self-knowledge distillation mechanism to distill a transductive graph neural network (teacher) to an inductive network (student). This transforms computationally expensive transductive learning into inductive learning, allowing us to use training patterns as the source of transduction and to inference in an inductive way. Specifically, we deploy the population graph as the transductive teacher model and the Multi-Layer Perceptron (MLP) as the student inductive network. MLP is a simple network that is easy to implement and can reduce

* corresponding authors.
This study is supported by grants from the National Natural Science Foundation of China (62276081,62106113), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (NO:ZYYCXTD-C-202004), The National Key Research and Development Program of China (2021YFC2501202), and the Major Key Project of PCL.

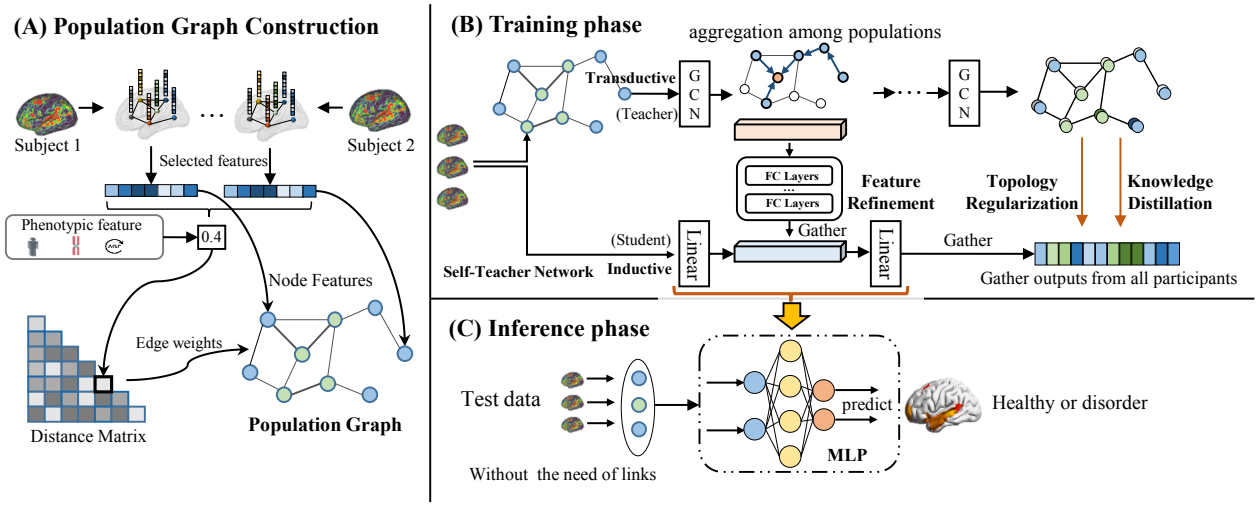

Fig. 1. Illustration of the population graph construction in (A), and the self-teacher architecture in (B)-training phase and (C)-inference phase.

computational costs by avoiding information aggregation from adjacent nodes. To preserve inter-subject associations from the population graph, we propose to regularize the student model training with topological structural knowledge of the teacher graph model. For evaluation, we implemented the ADNI dataset with 207 subjects with fMRI scans. Our contributions are detailed as follows:

- We propose a transductive-inductive network that leverages the advantages of both frameworks through self-knowledge distillation.
- We introduce the topology-regularized self-knowledge distillation (Topo-KD) method, which enables the transfer of inter-subject association knowledge.
- Evaluated experiments demonstrate that an inductive MLP network trained with Topo-KD could outperform a computationally expensive transductive GNN.

2. METHOD

The architecture consists of two parts: a transductive teacher network and an inductive student network. The teacher network is implemented using the population graph neural network, while the student classifier network is built using MLP. The training process is illustrated in Fig.1 (B), where self-knowledge distillation is employed to transfer graph refined features and topological knowledge to the student MLP network. During the inference phase, only the student MLP network is used for prediction, as shown in Fig.1 (C).

2.1. Population Graph and Graph Convolution

The population graph is constructed by representing each participant as a node. The feature and phenotypic similarity between participants are denoted as the edge weight, by reference to previous works [7] as: $e_{u,v} = corr(\hat{x}_u, \hat{x}_v) \cdot \sum_{h=1}^{H} \gamma(M_h(v), M_h(u))$, where H denotes a set of non-imaging phenotypic measures including sex, age, and site ID. The connectome similarity $corr$ is obtained by measuring the filtered features $\hat{x}$ by Recursive Feature Elimination as: $corr(\hat{x}_u, \hat{x}_v) = exp(-\frac{(\rho(\hat{x}_u)-\rho(\hat{x}_b))^2}{2\sigma^2})$ with a correlation distance ρ and a width of the kernel σ. γ is defined depending on the type of phenotypic value with a threshold θ: $\gamma(M_h(u), M_h(v)) = 1$ if $|M_h(u) - M_h(v)| < \theta$, else 0.

3-order Chebyshev spectral graph convolution is implemented to perform message passing among nodes as $g_\theta \star_G x = U g_\theta U^T x$, where U is the Fourier basis and $\star_G$ denotes a graph convolution operation on graph G. And $g_\theta(\Lambda)$ is obtained by $g_\theta(\Lambda) \approx \sum_{k=1}^{3} \theta_k T_k(\tilde{\Lambda})$, where $\tilde{\Lambda} = \frac{2}{\lambda_{max}}\Lambda - I_N$ with the identity matrix I_N, and maximum eigenvalues λ_{max}. $\Lambda \in R^{N \times N}$ is the diagonal matrix.

2.2. Topology-Regularized Self-Knowledge Distillation

Self-knowledge Distillation. Self-knowledge distillation reinforces the student network to learn from a weighted combination of the ground truth and the output distribution from the teacher with a weight λ: $L = L_{GT} + \lambda L_{KD}$, where L_{KD} denotes the Kullback-Leibler (KL) divergence, which is used to measure the distribution similarity between the teacher output p^T and the student output p^S as $L_{KD} = KL(p^T || p^S)$. L_{GT} is obtained by cross-entropy loss.

Feature Refinement. The intermediate representations

are predictive of the teacher network and are leveraged to reduce the gap between the teacher and student features. Considering that the latent features are in high dimensionality and there might exist heterogeneous relationships, we propose to introduce an auxiliary network to generate refined latent features. The auxiliary structure is denoted as ϕ and implemented with a 3-layer fully connected layer followed by ReLU activation and batch normalization layers. The objective function of the feature refinement is formulated as: $L_{refine} = ||\phi(h_l^S) - h_l^T||^2$. Due to the inconsistent sample size in transductive and inductive learning, all the training features and predictions are gathered and then distilled.

Topological Regularization. The population graph has benefited from predicting the nodes from features and their adjacent nodes with the links. Conventional self-knowledge distillation facilitates transferring knowledge from features but it fails to preserve the graph's topological structure, i.e., inter-node associations. In this regard, we regularize the student model training with the inter-node similarity with a weight $\psi_{u,v}$, which is obtained as:

$$L_{topo} = \sum_{u=1}^{N} \sum_{v \in N(u) \setminus u} \psi_{u,v} \cdot KL(p_v^T || p_u^S), \qquad \psi_{u,v} = \frac{e_{u,v}}{e_{max} - e_{min}} \gamma \tag{1}$$

The connection weight $\psi_{u,v}$ is obtained by normalizing the connection $e_{u,v}$. γ is to rescale the weight and is set as 1 for default. By this line, the prediction of sample u is attentively regularized by the transductive prediction of sample v with a weight $\psi_{u,v}$. For instance, a large edge weight $e_{u,v}$ indicates that the nodes are similar, and the $\psi_{u,v}$ approaches 1.

Optimization. In the training process, the objective function is constructed by a weighted combination of the target loss L_{GT}, knowledge distillation loss L_{KD}, intermediate feature refinement loss L_{refine} and the topology regularization loss L_{topo}, as displayed in Fig. 2. Moreover, to mitigate the cold start in the student model, we gradually balance the regularization terms with the weight $\alpha_t = \alpha_T \times \frac{t}{T}$, where T is the total epoch for training, α_T is set as 1.0. Finally, we obtain the loss function:

$$L = L_{GT} + \lambda_1 L_{KD} + \lambda_2 L_{refine} + \lambda_3 L_{topo} \tag{2}$$

3. EXPERIMENTS

3.1. Datasets and Image Processing

ADNI Dataset[1]: The ADNI dataset is a longitudinal multimodal neuroimaging dataset and is leveraged to predict MCI from AD. Notably, MCI is considered to be a significant stage

[1]http://www.adni-info.org/

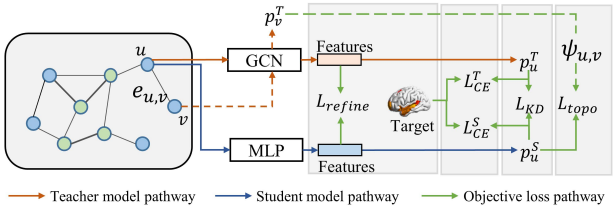

Fig. 2. Illustration of the network optimization. The solid line indicates the passing of the center node u, and the dotted line represents the passing of adjacent nodes, e.g. v.

for the preclinical diagnosis of AD. In this study, we collected 207 subjects for evaluation, comprising 103 participants with mild cognitive impairment (MCI) and 104 patients with Alzheimer's disease (AD). Only scans taken at baseline were included in the study.

All data underwent preprocessing with the Configurable Pipeline for the Analysis of Connectomes (CPAC) [12]. This included slice timing correction, motion realignment, intensity normalization, regression of nuisance signals, band-pass filtering (0.01-0.08 Hz), and registration into the standard space (MNI152). The regression of nuisance variables was modeled using 24 motion parameters.

3.2. Implementation details

In the implementation, the MLP network is a 3-layer multilayer perception classifier followed by a leaky ReLU activation. The number of selected features is 1000. λ_1 and λ_2 were set as 1, and λ_3 was set as 0.01. We set the learning rate to 3e-4, and the models were trained for 400 epochs. We employed 10-fold cross-validation, randomly selecting 10% of samples for testing in each fold. The model's performance was assessed by accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under the curve (AUC).

3.3. Competitive methods

In this study, these models are implemented and compared as baselines: (1) **SVM and MLP.** The support vector machine (SVM) and multi-layer perception (MLP) are implemented as baseline methods for classification. The upper matrix of the brain networks is fed into the classifiers for prediction. (2) **BrainNetCNN** [13] and **BrainGNN** [10]. BrainNetCNN and BrainGNN are two powerful approaches for brain network analysis. These two models are applied as inductive learning baselines for comparison. (3) **Hyper-GNN** [14] and **DHGNN** [15]. Hyper-GNN encodes the connectome into hyperedges, while DHGNN extends the Hyper-GNN into a dynamic graph. These models were implemented by reference to the originally proposed architecture. (4) **Population-GCN** [7], **TE-HI-GCN** [16]. These models were implemented using population graphs and compared as transductive learning

methods. Sex, age, and site are used to construct the graph.

4. RESULTS AND DISCUSSION

Classification Results. Table 1 displays the 10-fold cross-validation classification results of our proposed method and other competitive approaches, with the best results shown in bold. Compared with SVM and MLP, deep learning approaches that leverage graph structural information for classification showed improved performance. These methods can be divided into two categories: transductive learning (e.g. Population-GCN, TE-HI-GCN) and inductive learning (e.g. BrainNetCNN, and BrainGNN). In the ADNI dataset, transductive learning methods outperformed inductive learning methods in most cases. For example, Population-GCN and TE-HI-GCN achieved better accuracy (80.3% and 81.5%) compared to BrainNetCNN, BrainGNN, and Hyper-GNN (79.0%, 80.1%, and 78.1%). This indicates that graph neural networks with transductive learning perform better when dealing with a relatively small dataset by capturing associations among labeled and unlabeled data.

Compared to other methods, our proposed network demonstrated better performance, achieving an accuracy of 82.2%, sensitivity of 87.8%, specificity of 90.7%, and an AUC of 83.0%. We suspect that, on one hand, most existing methods suffer from over-fitting and poor generalizability. The improvements can be ascribed to the improved generalization performance raised by the knowledge distillation paradigm. On the other hand, our approach employs topology regularization and self-knowledge distillation, which penalizes the student prediction with topological information, refined features, and soft targets. This regularization paradigm has been shown to improve performance [17].

Table 1. The classification performance on the ADNI dataset across 10-folds. The best results are shown in bold.

Method	Acc	Sen	Spe	AUC
SVM	60.3±14.6	65.0±10.4	58.3±12.8	61.6±16.0
MLP	77.4±12.2	74.1±17.8	80.5±17.5	82.3±10.6
BrainNetCNN [13]	79.0±8.2	73.6±8.0	83.3±11.1	82.8±9.8
BrainGNN [10]	80.1±9.5	79.2±9.2	82.3±14.8	81.8±11.5
Hyper-Graph [14]	78.1±7.5	73.7±9.6	84.3±15.8	81.7±8.4
DHGNN [15]	81.4±7.8	84.8±11.9	86.4±12.1	81.5±6.4
Population-GCN [7]	80.3±11.5	85.3±11.3	83.1±16.3	81.9±13.7
TE-HI-GCN [16]	81.5±8.6	81.8±8.6	84.6±9.8	82.5±7.3
Ours	**82.2±4.9**	**87.8±8.2**	**90.7±10.2**	**83.0±6.7**

Table 2. Ablation studies on the effect of the feature refinement and the topology regularization on ADNI dataset.

Model	Components		ADNI	
	L_{refine}	L_{topo}	Acc	AUC
Teacher (GCN)			80.3±11.5	81.9±13.7
Student (MLP)			79.1±8.1	78.5±8.4
Student (MLP)	✓		79.9±7.5	79.2±9.7
Student (MLP)		✓	80.5±6.1	81.6±7.8
Student (MLP)	✓	✓	82.2±4.9	84.0±6.7

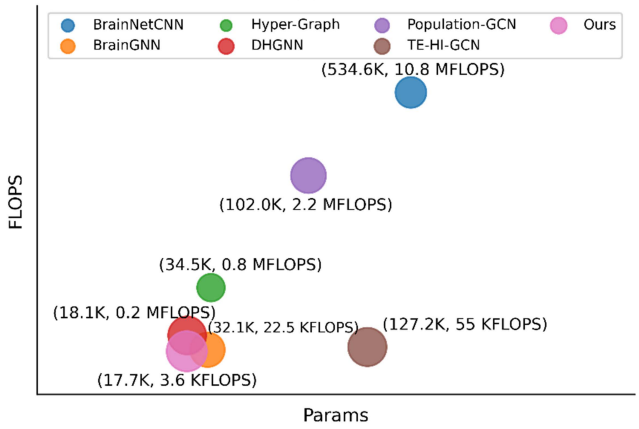

Fig. 3. Illustration of the parameters and FLOPS.

Ablation Studies. We conducted ablation studies to assess the impact of feature refinement and topology regularization on our model. Table 2 presents the results including the averaged accuracy and AUC across folds. We observed that the distilled MLP (second row) performed worse than the teacher population-GCN model (first row). However, with feature refinement and topology regularization, the performance on the ABIDE dataset surpassed that of the teacher model. This is consistent with previous studies that suggest the student model can outperform the teacher model [18].

We displayed the computational cost in Fig. 3, where the X-axis represents the number of parameters and the Y-axis represents floating point operations (FLOPS). The size of the circles represents the accuracy performance with larger circles indicating better performance. The preferred methods are those in the lower-left corner with low computation cost and parameters. For comparison with transductive and inductive frameworks, we accumulated all samples in each dataset to obtain FLOPS. The results show that Population-GCN has a relatively high number of parameters and FLOPS. Hyper-GNN and BrainGNN are relatively small in size. In comparison, the MLP trained by Topo-KD achieved the best accuracy with the lowest cost among all methods.

5. CONCLUSION

In this study, we propose a unified transductive-inductive framework for brain disorder diagnosis and introduce the Topo-KD approach to transfer topology-regularized knowledge to an MLP network. Our approach studies the properties of both transductive and inductive learning frameworks and transforms the transductive knowledge into an inductive learning method for inference. Our experimental results demonstrate the distilled inductive network enables generalization to new data and holds potential for large-scale inference and brain disorder diagnosis.

6. REFERENCES

[1] Edward T Bullmore and Danielle S Bassett, "Brain graphs: graphical models of the human brain connectome," *Annual review of clinical psychology*, vol. 7, pp. 113–140, 2011.

[2] Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, Alzheimer's Disease Neuroimaging Initiative, et al., "Benchmarking functional connectome-based predictive models for resting-state fmri," *NeuroImage*, vol. 192, pp. 115–134, 2019.

[3] Zarrar Shehzad, Clare Kelly, Philip T Reiss, R Cameron Craddock, John W Emerson, Katie McMahon, David A Copland, F Xavier Castellanos, and Michael P Milham, "A multivariate distance-based analytic framework for connectome-wide association studies," *Neuroimage*, vol. 93, pp. 74–94, 2014.

[4] Yanwu Yang, Chenfei Ye, Junyan Sun, Li Liang, Haiyan Lv, Linlin Gao, Jiliang Fang, Ting Ma, and Tao Wu, "Alteration of brain structural connectivity in progression of parkinson's disease: A connectome-wide network analysis," *NeuroImage: Clinical*, vol. 31, pp. 102715, 2021.

[5] Yanwu Yang, Chenfei Ye, Xutao Guo, Tao Wu, Yang Xiang, and Ting Ma, "Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning," *IEEE Transactions on Medical Imaging*, 2023.

[6] Yanwu Yang, Guoqing Cai, Chenfei Ye, Yang Xiang, and Ting Ma, "Tensor-based complex-valued graph neural network for dynamic coupling multimodal brain networks," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, vol. 48, pp. 117–130, 2018.

[8] Yanwu Yang, Chenfei Ye, and Ting Ma, "A deep connectome learning network using graph convolution for connectome-disease association study," *Neural Networks*, vol. 164, pp. 91–104, 2023.

[9] Giorgio Ciano, Alberto Rossi, Monica Bianchini, and Franco Scarselli, "On inductive–transductive learning with graph neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 758–769, 2021.

[10] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan, "Braingnn: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, vol. 74, pp. 102233, 2021.

[11] Yue Gao, Zizhao Zhang, Haojie Lin, Xibin Zhao, Shaoyi Du, and Changqing Zou, "Hypergraph learning: Methods and practices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2548–2566, 2020.

[12] Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S Ghosh, Chaogan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, et al., "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac)," *Front Neuroinform*, vol. 42, pp. 10–3389, 2013.

[13] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh, "Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.

[14] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao, "Hypergraph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 3558–3565.

[15] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao, "Dynamic hypergraph neural networks.," in *IJCAI*, 2019, pp. 2635–2641.

[16] Lanting Li, Hao Jiang, Guangqi Wen, Peng Cao, Mingyi Xu, Xiaoli Liu, Jinzhu Yang, and Osmar Zaiane, "Te-higcn: An ensemble of transfer hierarchical graph convolutional networks for disorder diagnosis," *Neuroinformatics*, pp. 1–23, 2021.

[17] Yanwu Yang, Guo Xutao, Chenfei Ye, Yang Xiang, and Ting Ma, "Regularizing brain age prediction via gated knowledge distillation," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 1430–1443.

[18] Yanwu Yang, Xutao Guo, Chenfei Ye, Yang Xiang, and Ting Ma, "Creg-kd: Model refinement via confidence regularized knowledge distillation for brain imaging," *Medical Image Analysis*, vol. 89, pp. 102916, 2023.